Measuring and Ensuring Excellence in Government Laboratories: Practices in the United States

Susan E. Cozzens[1]
School of Public Policy
Georgia Institute of Technology

with Barry Bozeman[2] and Edward A. Brown[3]

Prepared for the Canadian Council of Science and Technology Advisors
22 January 2001

Table of Contents

Executive Summary 2
Introduction 3
I. U.S. Federal Laboratories: Origins and Policy Studies 4
II. The Components of Excellence 6
III. Findings 9
IV. Case Studies
A. Agricultural Research Service, Department of Agriculture (ARS) 11
B. Army Research Laboratory, Department of Defense (ARL) 13
C. Department of Energy (DOE) 19
D. Environmental Protection Agency in-house laboratories (EPA) 22
E. National Institute of Standards and Technology (NIST) 24
F. National Institutes of Health intramural programs (NIH) 27
G. Naval Research Laboratory, Department of Defense (NRL) 29
List of interviews 32
Bibliography 32

---

[1] Atlanta GA 30332-0345, susan.cozzens@pubpolicy.gatech.edu
[2] School of Public Policy, Georgia Institute of Technology
[3] MITRE Corporation

Executive Summary

In the year 2000, the United States government obligated about $24 billion to government laboratories. These laboratories are enormously diverse, and serve many different kinds of government functions, from research and development through testing and evaluation. Mirroring their diversity in missions and organizational arrangements is an immense diversity in funding and management systems.

This report examines the practices U.S. government laboratories currently use to measure, ensure, and communicate excellence in their work. Rather than attempting to be comprehensive, the report illustrates the range of approaches with examples drawn from several government departments. Interviews with agency officials and publicly available documents form the basis for the description.

Excellence can be understood in four dimensions: relevance, quality, communication, and ethics. The report identifies best practices in each of these areas.

Relevance is coming to be better defined in many U.S. government agencies under the influence of the Government Performance and Results Act (GPRA), which requires strategic planning and mandates the alignment of agency activities with strategic goals and objectives that have been reviewed with stakeholders. Strategic plans are helping agencies set priorities, and setting the criteria for relevance evaluation by independent assessment panels. Agencies are increasingly involving users in both planning and assessment. In some laboratories, direct feedback from customers on a project-by-project basis is a useful supplement to program level relevance assessment.

Quality must be assessed by technically competent peers, in the view of U.S. program officials. Regular peer assessment of program quality is spreading as a practice in the mission agencies, and agencies are linking the protocols for such assessment panels to their strategic planning processes. Performance benchmarking, still generally done through qualitative judgment, is being incorporated in some cases into these protocols. Strong personnel evaluation systems provide the bedrock of quality control in every laboratory.

Communication of research itself and of its excellence are crucial processes that maintain the credibility of government laboratories. Where government laboratories produce results that feed into the regulatory process, strong data quality procedures need to be in place, and written products receive strong peer review. For communicating the excellence of research, reports of achievements still appear to be the most effective means. In some agencies, these reports are being standardized and evaluated for data quality under the requirements of GPRA.

Research *ethics* are strongly enforced in review of human subjects compliance by Institutional Review Boards and through investigation of scientific misconduct. Responsibility to the public is reinforced through outcomes-oriented strategic planning, as required under GPRA, while the ubiquitous application of critical assessment through peer review helps maintain the responsibility of the research community to truth.

In sum, the assessment of excellence in U.S. government laboratories is becoming more systematic, outcome-oriented, and linked to planning. These trends are to be expected under the accountability legislation currently in place.

Introduction

In the year 2000, the United States government obligated about $75 billion for research and development. A third of that amount or about $24 billion was obligated to government laboratories. About a third of the government laboratory amount went to laboratories that are operated by contractors for the government,[4] and two-thirds to laboratories operated by government itself.

These laboratories are enormously diverse, and serve many different kinds of government functions. They are also interwoven with each other, with one often serving as a contractor to another. Further, they are interwoven with industry, frequently subcontracting significant portions of their work to private firms. Mirroring their diversity in missions and organizational arrangements is an immense diversity in funding and management systems.

Because the labs vary so much, we will not attempt a comprehensive description in this report, either of the laboratories or how they are managed. Instead, we have chosen a few examples to illustrate the range. Our examples are drawn from the major agencies that house and manage government laboratories. While very different from each other in many ways, they also reveal common elements in laboratory management.

Table One
Participating Laboratories

Agricultural Research Service, Department of Agriculture (ARS)

Army Research Laboratory, Department of Defense (ARL)

Department of Energy multi-purpose laboratories (DOE)

Environmental Protection Agency in-house laboratories (EPA)

National Institute of Standards and Technology (NIST)

National Institutes of Health intramural programs (NIH)

Naval Research Laboratory, Department of Defense (NRL)

The questions motivating our analysis come from a report of the Canadian Council of Science and Technology Advisors, *Building Excellence in Science and Technology: The Federal Role in Performing Science and Technology* (BEST). BEST identified four major functions for the Canadian government laboratories:

- Support for decision making, policy development and regulations
- Development and management of standards

[4] Generally known as Federally Funded Research and Development Centers, or FFRDC's

- Support for public health, safety, environmental or defense needs
- Enabling economic and social development

The report saw three major challenges for the laboratories: an impending shortage in the human capital needed to fulfill the government's S&T roles; inflexibility in human resource practices and policies; and the aging and obsolescence of facilities, equipment, and research platforms. Among the principles BEST articulates that should guide the conduct of all Canadian federally performed and funded S&T is the concept of excellence, "meeting or exceeding international standards for scientific and technological excellence, and delivering social or industrial relevance, through openness, transparency, and regular and appropriate expert review." This report describes how U.S. federal laboratories assure and measure excellence.

The U.S. laboratories included in this study stretch across the four functions identified in BEST. The Environmental Protection Agency's laboratories primarily support the development of environmental standards and environmental decision making. The National Institute of Standards and Technology is the national metrology laboratory and a primary source for many U.S. manufacturing standards. NIH supports government's health mission, ARL and NRL support its defense mission, and ARS supports its agricultural mission. The Department of Energy laboratories stretch across functions, from defense and energy missions to basic research. Many of the laboratories work on environmental issues, which touch every area of government.

The report is organized in four sections. In the first section, we review the historical background of U.S. federal laboratories and the various policy studies that have addressed them. Here we also review the requirements introduced by the Government Performance and Results Act (GPRA), accountability legislation that has influenced planning and performance measurement in some of the laboratories. In the second section, we look across the illustrative laboratories with regard to four aspects of excellence: quality, relevance, communication, and ethics, identifying best practices. The third section summarizes the findings, and the fourth section provides a brief description of each laboratory or set of laboratories.

I. U.S. Federal Laboratories: Origins and Policy Studies[5]

The U.S. government has run its own laboratories for centuries. The U.S. Geological Survey has been in operation since 1879, and still provides geologic, topographic, and hydrologic information that contributes to the management of the nation's natural resources. The Agricultural Research Service was founded in 1934, to provide a scientific basis for improvement in productivity in what was then the major U.S. industry. The intramural laboratories of the National Institutes of Health were also first established in the 1930s.

The watershed for U.S. government laboratories, however, as for the federal role in the research system as a whole, was World War II. During the war, the military Office of Scientific Research and Development (OSRD) established a number of new weapons

[5] This section is based in part on Michael Crow and Barry Bozeman, *Limited by Design: R&D Laboratories in the U.S. Innovation System* (New York: Columbia University Press, 1998), especially pages 52-72 and Chapter 7.

laboratories around the country. After the war, the Atomic Energy Commission was given control of these laboratories, and transformed them into nine multi-program laboratories, including Argonne, Brookhaven, Sandia, and Los Alamos. Other laboratories within the military, with long and distinguished records, continued to thrive as the services became more and more technology-dependent. Sputnik was a second impetus to the growth of U.S. government laboratories. In response to this Soviet success, the U.S. constructed the National Aeronautics and Space Administration (NASA) from an older organization that had been developing aircraft for military use since 1915. The energy and environmental crises of the 1970s added two new agencies to the U.S. government, both of which included in-house laboratories -- the Department of Energy, which incorporated the AEC multi-program laboratories along with other activities, and the Environmental Protection Agency.

"With $25 billion on the table annually, and with facilities the size of small cities, it is little wonder that the U.S. has struggled with what to do with its federal laboratories."[6] Since 1978, more than 20 major commissions or task force groups have addressed issues in the shape, growth, and management of U.S. government laboratories. In 1978, a report on the DOE laboratories stressed organizing the labs for a national, rather than programmatic, focus, and coordinating missions to maximize individual and laboratory group competencies. In 1983, a federal laboratory review panel chaired by David Packard stressed clarifying mission, letting function determine size, and developing a personnel system independent of the civil service. In 1993, a report on defense conversion from the Office of Technology Assessment recommended speeding up Cooperative Research and Development Agreements (CRADAs), enhancing local autonomy for laboratories, and creating a laboratory rationalization commission. Another 1993 report, by the Council on Competitiveness, a group with strong industrial representation, stressed industry as a customer of the federal laboratories. It made recommendations for the technology transfer process, including increasing the discretionary budgets of laboratory directors to 5-10% of their annual budgets and assigning 10% of the budgets of DOE and NASA laboratories to technology transfer programs. The Galvin Commission in 1994 recommended new management systems based on quality principles. They urged laboratories to develop clear missions and emphasize the use of federal laboratory core competencies in collaboration with the private sector and other federal agencies. The Commission also recommended focusing the laboratories on long-term R&D, as part of the national innovation system.

Another element that has entered the environment for U.S. government laboratories in the 1990s is the Government Performance and Results Act of 1993 (GPRA). GPRA requires all agencies to prepare strategic plans (every three years), performance plans with target levels of performance on specific objectives (every year), and performance reports (every year) that say whether the objectives have been met. Some of the illustrative laboratories in this study have been significantly affected by their response to GPRA; others have barely been touched. ARS and EPA have worked hard to use GPRA to help organize and strengthen their laboratory systems. The Defense Department laboratories are so far down in the organization that they have not been required to do so, although the Army Research Laboratory served as a pilot project under

---

[6] Crow and Bozeman, ibid, p. 215.

GPRA. NIST was already working hard on performance metrics before GPRA was passed, and has spent somewhat more resources on those efforts since that time. The NIH intramural laboratories are following the lead of their agency, which does not find the GPRA framework useful for management.

II. The Components of Excellence

Relevance. In the illustrative laboratories visited in this study, GPRA has had its clearest effects in the area of alignment with government missions. Both ARS and EPA have been working hard under GPRA to create an organization that shows how their activities are in line with the agency's newly articulated mission and goals. This is a first step in re-organizing and redirecting toward those goals. Both agencies have developed national program structures that group the activities of individual laboratories (which are scattered around the country) into coherent themes that align with the agency's strategic plan and objectives. These processes have helped the agency examine each piece of its activity anew for its importance with regard to mission, and more clearly articulated the criteria for making that judgment. Over the long haul, this process will affect budget allocations. EPA's strategy has tightened considerably over the last five years, in response both to GPRA and to other external pressures. Projects must not just be relevant: they must also hold promise of order of magnitude improvements in risk management strategies or significant changes in risk assessments. ARS reports that its new national program structure improves communication both internally and externally about what it is producing for the public.

The other laboratories have been influenced less by GPRA, but all clearly incorporate processes for aligning their activities with government missions and goals. The two defense laboratories use strong customer relations for this purpose. Both ARL and NRL organize their work in projects. In NRL's case, other services or government agencies fund 80 percent of its projects. Considerable management attention goes into the process of alignment. Senior managers keep track of the research directions that might contribute to client goals, and set directions for the laboratory that positions it to help clients best. ARL has its own funding, but has signed agreements with its customer organizations to provide the research they need. At the end of each project, ARL sends a simple questionnaire asking that customer about satisfaction. It has also established an advisory board that consists of its Army clients, and both laboratory and clients are pleased with its operation. NIST and NIH also pay attention to the stakeholder base for each of their programs, and incorporate relevance judgments into regular program review processes.

The information shared through linkages with other parts of government plays an important part in the process of assuring relevance. The laboratories vary quite widely in the extent to which they stress linkages either within or outside government in their operations. NRL represents the extreme case. It would not survive without strong links to the other services and to other government agencies. Likewise at NIST, a significant portion of the work is supported by other government agencies in support of their missions. GPRA seems to have encouraged the linkage process at ARS and EPA. Both agencies report more internal coordination, more consultation with other parts of the larger agency and between widely scattered laboratories. NIH reports the lowest levels of attention in this regard. National Advisory Committees represent external constituencies

in the NIH structure, but directors of the intramural programs report to them on a courtesy basis. The Boards of Scientific Counselors that review intramural programs at NIH are less likely to include members that clearly represent constituencies. Much linkage probably goes on at individual laboratory and branch level, but encouragement for it is not prominently displayed in the evaluation criteria used to review those units.

In this area, two indicators seem to be used commonly among the agencies. First, the existence and continuation of funding from other parts of government or from external customers provides one measure of success. As the NRL Director put it, "The best metric of performance for our divisions is whether they get more work from their sponsors." Second, several laboratories track CRADAs (Cooperative Research and Development Agreements), a mechanism invented in the late 1980s to allow private organizations access to government expertise. In the early 1990s, CRADA counts were becoming a central performance metric for government laboratories. Concern arose that that the count was becoming more important than the relationships themselves. A reaction then set in that put this metric back in perspective as only one among many that indicate laboratory linkages.

Quality. Regardless of their other management processes, each of the laboratories reviews its programs and people for scientific quality using external reviewers with high scientific credibility. Formal, quantitative benchmarking of the quality of laboratory work is seldom done. Instead, the external reviewers are the calibrating instrument. Laboratories count on them to provide a certification of quality if warranted or if not, an early warning of problems.

At program level, several of the review processes for these laboratories have been in place for decades. NIST has been using a part of the National Research Council[7] to assess its programs since the 1950s, and the NIH system of Boards of Scientific Counselors has been in operation at least as long. Several years ago, ARL set up an external review process patterned on NIST's and also housed in the National Research Council. ARS has just set up a program review process as well, and it has recently completed its first review. The new system at ARS is strongly geared to judging programs by GPRA goals and objectives, and has been designed to allow projects to be seen in their program context for the first time. NRL uses program review for only part of its activities, the Base Program, which is funded by the Office of Naval Research. ONR mandates these reviews for all the programs it funds, and uses them as an upward accountability mechanism within the service.

Personnel review is detached from these program review processes in most of the laboratories. Standard annual performance reviews of course take place in all of them. At ARL, key performance metrics are built into the performance plans of managers. NIH has a strong system of external review for promotion and tenure, modeled on university systems. Since one of the few routes to program change at NIH is personnel turnover, the tenure and review process is taken very seriously. ARS uses an internal group to do periodic reviews for promotion.

---

[7] In the U.S., this is a part of the National Academies complex, a quasi-governmental organization that performs a great deal of analysis and evaluation for government.

The personnel inflexibility that government laboratories face around the world also characterizes most of the U.S. laboratories. One agency pointed out that it is virtually impossible to close a laboratory, because the Congressional delegation from that area would intervene to save it, regardless of its value to management. NRL's successful approach to personnel flexibility therefore stands out. NRL does not have its own appropriation and operates entirely through projects funded by sources outside the laboratory. That funding stream provides a clear criterion for when to let a government researcher go. If the external customer base is no longer there, no money is available to pay the person, and the RIF process (reduction in force) begins. As many as three percent of NRL's employees may depart in any given year through this route.

Communication. GPRA and GPRA-like processes in U.S. federal laboratories has significantly improved the transparency of the work being performed there, by involving stakeholders in the planning and evaluation process and by creating program structures that are more visibly aligned to public purposes. The Agricultural Research Service provides an example. GPRA has inspired ARS to provide an annual report on each of its national programs on its web site, with links to projects in specific states. Workshops with customers and stakeholders are used to develop action plans, which get public exposure through the web site and mailings. Program staff are delighted with the new structures, since others now understand better why they do what they do. The Army Research Laboratory assessment system also includes communication with stakeholders as a key element. Their Stakeholder Advisory Board consists of high-ranking officers from the services ARL supports. This group receives information about the laboratory in an annual visit, and reviews other assessments for relevance, to give advice to the ARL Director. The Naval Research Laboratory uses direct interaction with its various funders as the major channel for communicating findings.

NIH addresses the dual problems of communicating the quality and relevance of its research with the use of three special kinds of reports on achievements. Science Advances (one page) and Science Capsules (one paragraph) report on a specific scientific discovery published during the past year and supported by NIH funding, including its significance for science, health, or the economy. Stories of Discovery focus on a topic and trace its major development over several decades, connecting those advances with improvements in the quality of life, health, and health care, as well as economic benefits. Audiences for NIH performance reports have warmly received these disciplined reports of accomplishments.

Several of the agencies need to communicate the findings of government science in controversial areas. The Department of Energy balances security restrictions (in those laboratories that serve military functions) with the need for public accountability, particularly in its hazardous waste cleanup efforts. DOE uses face to face approaches, like town meetings and workshops, to consult with local communities around its sites. The Environmental Protection Agency uses extensive peer review of all documents to be released to the public to help protect its credibility in the highly-charged arena of environmental regulation.

Ethics. Public responsibility is an inherent element of the research assessment process. In the United States, the phrase "research ethics" usually refers to specific issues that arise in the research process. Protection of human subjects is one of these. Strong

regulations exist to ensure this protection, and both government laboratories and academic institutions are required to have Institutional Review Boards that review any research protocol involving people for compliance. This is taken very seriously. In addition, enforcement of regulations preventing fraud, falsification, and plagiarism has strengthened over the last decade, in response to several well-publicized scandals in the late 1980s. Again, institutions have primary responsibility. Research ethics handbooks outlining responsibilities with regard to sharing of data and credit have been prepared.

Beyond these specific aspects of the research process, however, lie several larger responsibilities that receive less attention under the term "ethics" but are also taken quite seriously in U.S. government-sponsored research, either in government laboratories or in universities. First is the responsibility to put the public first in whatever research is undertaken. Some feel that this ethic has been undermined in U.S. academe by the (government-sponsored) emphasis on research with economic potential. Many see the relative independence of government laboratories from that pressure as a strength. Public accountability in this sense is strengthened through the GPRA process, with its emphasis on outcomes for the public, and through relevance review.

The ultimate responsibility of a researcher, of course, is to the truth. Peer review is the process that teaches and enforces that value. The program review processes described throughout this report are thus one means by which this value is maintained. Another is the especially stringent peer review process applied to government laboratory reports that have regulatory implications. The description of the Environmental Protection Agency in Section IV illustrates.

III. Findings

The cross-cutting themes of the last section help us to identify the prerequisites for excellence in government research in the United States. Clear goals are needed to align laboratory efforts with the expectations of external groups. Adequate resources must be provided, or the laboratory does not have the technical capacity to achieve its objectives. Serious, regular review processes by external experts appear to be particularly necessary in government labs, which lack the competition for grants that keeps university researchers in the U.S. at their best. The flexibility to start and stop lines of research, laboratories, and personnel is crucial.

The impediments to excellence in U.S. government laboratories follow a similar pattern. Large agencies that have grown through the accretion of legislation often have conflicting or competing legislated missions, that cannot be sorted out by even the best of strategic planning processes. Personnel and procurement policies often do not allow government laboratories to buy state of the art equipment or compete with academic and industrial ones for the best people. Politics can intrude into the rational ordering of laboratory life, creating challenges to methods or roadblocks to good management.

Within this set of opportunities and constraints, a few best practices in ensuring excellence are emerging among the agencies that manage U.S. federal laboratories. At the stage of program definition, which sets the criteria for project selection, stakeholder-based strategic planning is not only required by GPRA but is also being adopted as a best practice. This process insures the alignment of agency efforts with the expectations of relevant parties in its environment, whether they be customers, potential users of research

results, the Office of Management and Budget, or Congress. While developing such a strategic plan is expensive in staff time, those agencies that have produced one (for example, the Agricultural Research Service) have been happy with the results. The strategic plan itself can be an important communication mechanism, both outside the laboratory and within, and it sets clear criteria for shifting funds when needed. It thus makes management more effective.

At the stage of the conduct of research and examination of its results and impacts, best practice in government laboratories in the United States is a regular process of program review by an independent expert review panel. When agencies are challenged by outside criticism, they strengthen these processes in the directions of greater independence, transparency, and consistency.

The NIST and ARL processes are good examples of such processes, in large part because an external body with a strong reputation for independence and high technical credibility handles them. Such panels always include active researchers in the technical field being examined, and for any review that does not have the research community itself as its primary customer, users are also included. The users are also people with technical backgrounds, but who bring applications knowledge as well. The review panel receives information on the program, usually including personnel and equipment data; a list of publications or other outputs; and the program's own reports of progress, including any special achievements or awards. For government laboratories, site visits with presentations by research staff are standard. The work of each laboratory is ongoing, so review on a regular schedule (three years is usual) examines a stream of results in relation to the goals and objectives articulated in the strategic planning process. This simple pattern can be applied to a wide range of programs.

Panels like these judge government research on its relevance to mission (using the strategic plan), its quality, and often a comparison with similar work in other settings (benchmarking). In general, none of these criteria is quantified. Asking questions that are too specific tends to insult high-level reviewers, and is generally avoided. Reviewers put specific technical advice to management in their own words, and alert management to any serious problems with personnel or projects. Their reports are generally shared with program staff, who have a chance to reply, and often have to report back at a later time on how they have addressed any issues raised.

Including criticism in the reports of course makes public reporting problematic sometimes. Some U.S. agencies keep their review reports internal, but those that make them public are more pleased with their results. GPRA places a special value on making such reports public if they are used in performance reporting, and agencies that have been less than open about who has done the review or what they said have been criticized by Congressional staff.

Such regular panel reviews are costly, in staff preparation time, in travel costs for panel members, in the time of panel members, and in administrative costs at outside bodies when they are used. The historical record of such processes, however, is that they identify serious problems with 3-5% of the portfolio examined. When the agency fixes these problems, it is rewarded for the substantial investments in the review process with substantial improvements in performance.

IV. Case Studies

A. Agricultural Research Service

The Agricultural Research Service (ARS) is the principal in-house research laboratory of the U.S. Department of Agriculture. Its budget is about $745 million, which is spread among about 1100 projects grouped into about 22 national programs. About 2000 scientists are employed, in facilities located in about 100 locations around the country. The Department also supports state-level agricultural experiment stations and houses a small program of extramural grants.

ARS has revamped its program structures and assessment processes significantly recently. One influence is GPRA, but another is criticism the Service received in 1998 in a study by the National Academy of Sciences. The national program structure is new since then, and a new program review process was implemented in 2000.

Program structure. The ARS strategic plan has five goals, which translate the goals of the Department into research terms.

- Through research and education, empower the agricultural system with knowledge that will improve competitiveness in domestic production, processing, and marketing.
- To ensure an adequate food supply and improve detection, surveillance, prevention, and education programs for the American public's health, safety, and well-being.
- A healthy and well-nourished population who has knowledge, desire, and means to make health-promoting choices.
- To enhance the quality of the environment through better understanding of and building on agriculture's and forestry's complex links with soil, water, air, and biotic resources.
- Empower people and communities, through research-based information and education, to address the economic and social challenges of our youth, families, and communities.

Over the last few years, ARS's projects have been grouped into "national programs" that are aligned with the goals. Working groups of the National Program Staff have done this work, whittling an initial list of about 50 program areas down to the current 22 or so. Examples are "crop production" and "integrated crop systems." These national program areas represent a major cultural change for the organization. The programs set expectations for specific projects. Under this structure, every laboratory activity is linked to the strategic plan goals. The strategic plan forms the criteria for placing projects into programs, and will form the criteria for assessing them in each year's budget submission. Thus, in the part of the process, relevance to Department goals is a dominant consideration. Another feature of the new national program structure is the appointment of 30 national program leaders. Each has an area of special expertise, and is responsible for coordination in that area.

Project assessment. A new Office of Scientific Quality Reviews for ARS has been set up in the National Program Staff. Under this system, all the projects in a particular program will undergo external review at the same time. This will allow panels to see the overall structure of the program, rather than seeing just one project at a time. The first review under this system, of Food Safety, has just been completed. The external panels make decisions on individual projects (rating them as outstanding, good, or bad). They will judge these units by quality and relevance, plus the capability of the proposers to do the project. They will recommend changes as needed. Most of the members of these panels work outside ARS, and most work outside government. All are Ph.D. scientists with excellent research credentials. Some combine research expertise with experience in organizations, like seed firms, that use ARS research results.

ARS feels that this new system has many strengths. In the past, the national program staff organized review of each project when it expired, requesting three to five external reviews. Reviewers were not convened, and they were not paid. The new review system allows for more of an overview of the program, and they feel that convening the panels allows them to understand program balance and generate better advice. The first review panels have been very rigorous, and ARS staff are pleased with the input they are getting.

Assessment in the research process. Between program reviews, managers in the laboratory are responsible for maintaining progress toward project goals. Each area office reviews its own activities annually, and submits a report on accomplishments and impacts.

In addition, the 2000 scientists of the ARS are also subject to rigorous individual review processes. Every three to five years, an internal peer review group examines their accomplishments, to determine whether they should be promoted or remain in grade. This process has great credibility in the agency, which puts considerable time and attention into it.

Results and impacts. The program formation process already described takes projected results and impacts into account, as do the prospective program reviews. A retrospective review process is under consideration. It will use the goals and objectives of the ARS and program area strategic plans as criteria.

Communication. The new structure of national programs is designed to communicate the benefits of ARS research to the public. By bringing projects together under themes that are clearly tied to the strategic plan, ARS hopes to make its usefulness evident. Its home page has been organized to give public access to this information. An annual report for each national program area appears, with links to a map that shows the projects associated with each. In addition, ARS holds occasional workshops to consult with customers and stakeholders in specific program areas, developing action plans that are also displayed on the Web. Results are also mailed.

An important communication benefit of the new structure has been with the scientific staff. According to agency officials, they used to wonder whether anyone was paying attention. Now they know that someone does. ARS researchers are thinking more about their impacts now, and are aware of the need to solve problems that customers and stakeholders want solved. In addition, according the agency's budget staff, the questions

they are receiving from Congressional staff have changed, to focus more on outcomes for the public.

B. Army Research Laboratory prepared by Edward A. Brown

The Army Research Laboratory (ARL) is the U.S. Army's corporate laboratory. It provides the technological underpinning for the Army's acquisition programs and is the lead organization in developing the technology necessary to fulfill the Chief of Staff's vision to transform the Army. ARL is a 2200-person organization with approximately 1300 scientists and engineers. Its annual budget is approximately $670M. It is located at two main locations, the Adelphi Laboratory Center and Aberdeen Proving Ground, both in Maryland and both with about 850 people, five other sites around the country with between 50 and 150 people each, and then at more than two dozen other sites, both in the U.S. and abroad with just a few people at each.

ARL was activated in 1992 as a consolidation of seven formerly independent Army laboratories, the purpose being to form a central organization that would be responsible for most of the basic and applied research carried out by or for the Army. ARL has no systems development responsibilities other than to provide technical support and consultation to those organizations that do. In 1994 ARL volunteered to be one of about 80 agency pilot projects under the Government Performance and Results Act (GPRA) which required all agencies of the US Government to engage in strategic, long term, and annual program planning and reporting, and in performance evaluation, with the emphasis on results as opposed to former management schemes which tended to highlight inputs (such as expenditure of resources) and outputs (such as measures of activity). The enactment of this law caused obvious consternation among the various federal R&D organizations. ARL was the only R&D organization among the 80 pilots that took on the challenge to devise planning and evaluation methods that were not inimical to the research process. Over the course of the five year pilot program, ARL broke much new ground in these areas. The information provided below is based on these initiatives and innovations.

Scientific Program Decision and Proposal/Project Selection. For a corporate laboratory the decision concerning what to work on, both at the program (strategic) level, and the project/task level, is driven by several things. First and foremost is the organization's mission statement. This statement must flow from the mission statements of the parent command and that must flow from the Army mission. Thus, we see the requirement for linkage of missions to drive program formulation. From the ARL mission we derived several major strategic thrusts or vectors that the laboratory will move along for the foreseeable future (~5 years). For instance, in the Chief's current vision to transform the Army into a lighter, more mobile, more easily sustainable force, the requirement for lighter combat vehicles instantly becomes apparent as a major technology challenge. From the realization of this challenge one can immediately derive a whole set of long term programs/goals (lighter armor materials and structures, more fuel-efficient power trains, etc.) from which flow a host of near term (annual) projects. Some of these goals can be described in terms of quantifiable outputs and possibly in some cases, outcomes.

However, the nature of research being what it is, for the most part we are restricted to qualitative goals.

Selection among these projects is primarily limited by the available resources (funds and people). However, within this limitation choosing among the candidate projects brings into play several other concepts. The first is the requirement for a corporate lab (any corporate lab, public or private) to have a balanced program. The balance must be struck between the long-term, visionary, paradigm-breaking, opportunity-driven work, and the short-term, application-oriented, customer-driven work. Too much short term work and the lab soon loses its technological edge and becomes useless over the course of a few years; too much long term work and it is seen as a "hobby shop" and loses the support of its stakeholders. How this balance was achieved at ARL was by a construct devised at the lab's inception. Recognizing the potential problem of attaining an appropriate programmatic "balance", ARL entered into memoranda of agreements (MOAs) with each of its primary customer organizations, the Research, Development & Engineering Centers (RDECs) of the Army's commodity commands. In these MOAs ARL promised to expend "at least 50% of its mission program" (i.e. its direct funded program) on work for the RDECs. That is, this would be a "free good" to the customers. What precisely would be done was arranged annually through a process of constructing annexes to the MOAs called Technical Planning Annexes (TPAs) which were negotiated between the first level supervisors (e.g. Branch Chiefs) at ARL and at the customer organizations, and then signed by the ARL and RDEC directors. These TPAs spelled out the scope of work, what would be delivered and by when and to whom, and how much ARL would expend of its resources on each task. Thus, the customer was involved in the project selection process from the beginning.

For the other "50%", the long term, more speculative research, the Director used his discretion supported by advisory information from, among others, the ARL Fellows, a cadre of the most senior and highly honored researchers at the lab.

Scientific Inquiry. Determining whether there is an appropriate amount and kind of inquiry taking place at ARL is coupled to the larger question of how to determine the technical quality of the program. Programmatic advice can be found on every street corner in Washington, but sound, objective, unbiased technical evaluation is very difficult to obtain. To solve this problem ARL turned to the National Research Council, the operating arm of the National Academies of Science and Engineering. They contracted with the NRC to provide a peer review group comprising world class scientists and engineers who would come into the lab and perform an on-site review of programs and people looking specifically at the technical content. This Technical Assessment Board (TAB) consists of about fifteen individuals. Reporting to the Board are six panels of about ten people each, also of world class reputation. The panels, one in each of ARL's six business areas, visit the labs annually, take briefings, interact with the staff, examine the facilities, and then write a report to the Board. The Board combines these six reports into an overall assessment of the laboratory which is published by the National Academy Press and released as a public document. The report contains suggested action items for improvement and the public nature of the report (especially in a military environment) puts considerable pressure on the lab to act on these suggestions.

A concrete example of the process may help to clarify. In 1998, the TAB consisted of fourteen members, seven from academe, four from industry, one from a government laboratory, and two consultants. Subpanels assessed particular laboratories. The names and credentials of all TAB and subpanel members appeared in its report. The Board used four criteria for its review: formulation of the project's goals, connections to the broader community, technical methodology, and capacity for research. The TAB noted four cross-cutting issues: impressive progress in recent years, the need to refresh in-house expertise, the success of a federated laboratory program, and the need to develop "thoughtful, concrete" work plans.

Each of the panels met once to receive two days of unclassified briefings for staff, then visited laboratories, where about a third of the projects received intensive attention. The group reviewing the Survivability and Lethality Analysis Division (SLAD), for example, went to Aberdeen, Maryland (away from the main ARL site). The SLAD panel consisted of eleven members, four from industry and one with significant government laboratory experience. They focused on two areas within the division, a virtual test environment that was under development and the Nuclear, Biological, and Chemical (NBC) Branch. The panel devoted most of its attention to the first project, assessing its potential and suggesting ways to develop it fruitfully. The NBC work received a less glowing report. "The future for NBC activities... looks bleak," they reported. NBC funds were set to expire at the end of following year. "If the NBC Effects Branch continues within ARL," they wrote, "the SLAD analysts need an opportunity to develop better computational tools and gain expertise in the relevant physics, chemistry, and materials science." The panel identified a number of vulnerabilities in NBC analysis resulting from military application of civilian software, and noted places where the current models might be producing inaccurate results. For example, "The requirements for flat terrain and a simple wind field mean that the model is likely to predict higher levels of exposure to chemical and biological agents directly downwind of release than might actually be encountered."[8]

Results and Impacts. When ARL took on the challenges posed by GPRA whose principal focus was on outcomes, it became immediately apparent that it simply was not possible to discuss outcomes of basic research in any sort of meaningful way. Such outcomes were usually decades away and, indeed, might not even be definable in the present. This having been said and internalized by both ARL and the rest of the federal R&D community (and eventually by OMB, and the Congress), the question ARL posed to itself was what could reasonably be reported to its stakeholders to assure them that ARL was performing at a "world class" level? Three areas emerged to focus on. ARL could report on the relevance of its work to its customers, whether it was being productive (i.e., putting product out the door in a timely manner), and whether it was doing "world class" quality research. Answering these questions was as responsive to the requirements of GPRA as a laboratory could be.

Responding to the questions of relevance and productivity was the TPA process discussed above. In order to "measure" how well the lab was doing in providing the customers with the high quality products they needed in a timely fashion, ARL instituted

---

[8] Army Research Laboratory Technical Assessment Board. 1998 Assessment of the Army Research Laboratory. (Washington, DC: National Academy Press, 1999)

a targeted survey process. Simple straightforward surveys with a half dozen questions marked on a 1 to 5 scale were sent to the customer signatories on all the TPAs, as well as the customers for the reimbursable program, for a total of over 450 surveys a year. The survey forms were individually identified for each specific task. The surveys had over a 50% return rate.

The statistics from these surveys, along with comments from a comment box on the form were analyzed and reported as a measure of customer satisfaction in the relevance and productivity domains. Then at the end of the year there was an annual meeting of all the RDEC directors (the ARL Board of Directors – BOD) to validate that ARL had indeed committed the appropriate portion of its program budget to their benefit and that the survey results reflected their degree of satisfaction with the work product. As an additional incentive to satisfy the customer, for any individual score of 1 or 2 (poor or fair) on any survey form, or any negative comment, the SES-level ARL directorate head was responsible for contacting the individual who submitted the survey form within five working days to inquire as to the nature of the problem causing the low score and to take steps to correct it. This response system, along with a goal for the directorate's aggregate score was placed in each directorate head's performance standards.

Communication of Results and Impacts. Communication with ARL's stakeholders presented several problems, not the least of which was identifying who those stakeholders were. According to an MIT Sloan School study that ARL participated in, there are three groups of stakeholders for an organization's research enterprise: the set of immediate customers, the end item user, and the senior leadership of the firm. For ARL the immediate customers were the RDECs (and other reimbursable customers) as discussed above, and communicating them was done formally through the survey process and at the annual BOD meetings. There was also a considerable amount of informal communication carried on throughout the year. As a matter of fact, the most common negative comment on the survey forms occurred when the ARL project leader failed to talk to his customer frequently enough.

For the two other stakeholder segments, the end item user, which corresponded to the fielded troops and the senior leadership of the Army, it was readily apparent that a survey process was inappropriate. ARL conceived of a Stakeholders' Advisory Board (SAB) which was chaired by the Commanding General (4-star) of the laboratory's parent command, and comprising about a dozen of the Army's senior staff members (3-star or civilian equivalent level). This group would assemble annually at ARL. They would receive a "state of the lab" presentation from the ARL Director, hear some high level technical talks of current interest, and tour the lab facilities to see some real-time demonstrations of ongoing work. In order to pull together the other phases of this overall evaluation process, the SAB would also receive a report from both the TAB and the BOD. This would allow the SAB members to gain a total, strategic-level view of the lab's performance over the past year. The day would close with a round table discussion among the members assessing their findings and instructing the Director on their expectations for the coming year. This also included some number of specific action items that were recorded in the minutes of the meeting and were required to be completed and reported on before the next annual meeting.

The SAB process was found to be extremely effective as a way to accomplish several things, not the least of which was to communicate to these senior leaders how the lab was performing. This is usually difficult to do with individuals who are not technically trained or whose positions are not close to the R&D operations of the organization. It also provided a venue for ARL to approach these individuals with issues that needed high level attention for resolution. In addition, it enabled issues that might put ARL in conflict between two competing members of the Army leadership to be discussed and resolved in open forum between these competing interests, thus taking ARL out of the middle. Finally, it gave these members of the Army staff a feeling of ownership in their corporate laboratory that served ARL well throughout the year. There have been five annual SAB meetings to date, each one more successful than the other. The senior leadership of the Army has been extremely receptive and appreciative of the opportunity to be heard in such a forum.

Metrics. While metrics are usually the first thing that one thinks of when discussing performance evaluation, in the case of R&D ARL decided that metrics did not represent a true and useful evaluation which is why it adopted the processes described above. However, as the construct evolved ARL came to realize that while metrics cannot adequately describe the technical quality or relevance of the program, there was a very important fourth question that was meaningful to evaluate: How was the functional health and the research environment of the laboratory? Here we found many dozens of different metrics that served as useful indicators of the organization's health. The assumption was that while no number or set of numbers could in and of themselves guarantee that good science was being done, if the values of these metrics were all high (or low if that was the "better" value), then there was fertile ground in which good science could be done. If, on the other hand, the metrics were "bad", then it was certain that good science would not be done.

Implementing a system of metrics was done along two tracks. There were a collection of metrics indicative of the functional health of the lab that were automatically tracked by the fiscal and personnel systems such as obligation and disbursement rates, overhead rate, average age, grade, etc. of the work force, personnel turnover rate, and so on. These were tracked by the respective functional offices and only reported to the Director if they fell outside of appropriate bounds. Then there was a smaller collection of metrics which the Director determined would give him indications of the research environment such as papers and patents, percentage of doctoral level staff members, numbers of visiting scientists, etc. Here goals were set, partly by using the results of benchmarking peer organizations, both public and private, and partly by the Director's own intuition. Lab-wide goals were broken down by directorate and placed in the directorate heads' performance standards. It is important to understand that this metrics program was carried out mainly for internal management purposes. They were not briefed to stakeholders as part of any sort of regular process. The use of metrics in R&D is much too easy to subvert to punitive uses, so it has been ARL's policy to publish such metrics only where and when the Director decides is appropriate, and then only with sufficient caveats as to how to interpret them.

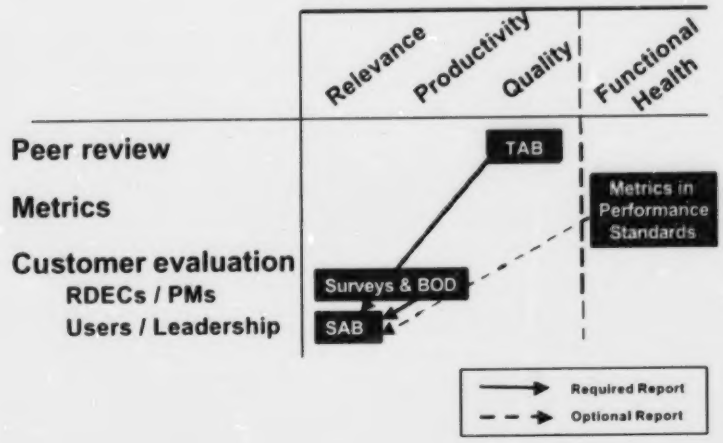
Summary. The collection of processes developed by ARL under the GPRA is known as the ARL Performance Evaluation Construct. Its implementation as described above is summarized in the following figure. Its strengths are many:

- It provides a comprehensive view of the R&D organization's performance,
- Through the SAB the various pieces are able to be tied together in a holistic form,
- It is quantitative where quantification makes sense for a laboratory, but it is qualitative where it needs to be,
- It speaks directly to each of the stakeholder/customer groups, and
- It is "non-threatening" from the standpoint of not being a numerical report card.

Its principal weakness is that it is costly to operate both in terms of funds and manpower. The contract with the NRC for the TAB is quite significant. The labor that goes into all the various parts of the construct is a major consumer of time, from the preparations for the SAB, TAB and BOD meetings to the labor involved in the survey and metrics processes.

However, all that being said, it has been ARL's experience that the effort has been well worth it. Not only has the lab generated enormous amounts of good will among its three stakeholder segments, but it has also provided a way for the lab to track improvements and report in this mixed qualitative/quantitative mode how well the laboratory has done. The bottom line is that the lab has shown continuous and dramatic improvement over the six years that the Construct has been in place.

Implementation of the
Performance Evaluation Construct



C. Department of Energy prepared by Barry Bozeman

The U.S. Department of Energy was created in 1977, but has roots extending from the 1942 beginnings of the Manhattan Project of the U.S. Army. After World War II, the U.S. Congress created the Atomic Energy Commission, chiefly aimed at finding peacetime uses of nuclear energy and regulating nuclear energy. The Department of Energy, created in 1977, brought together the functions of the Energy Research and Development Administration, the Federal Energy Administration, the Federal Power Commission and some of the energy science programs of the National Science Foundation. Currently, the Department includes approximately 16,000 employees and more than 100,000 contractor employees working in 50 major installations. Of the 15 Cabinet Departments, the Department of Energy is the second smallest in terms of personnel (if contractor employees are not included).

Much of the Department of Energy has little to do with research and, instead, focuses on such issues as hazardous waste remediation, nuclear security, and promoting conservation. It is a significant manager of a variety of facilities and infrastructures, having 2.4 million acres of land and more than 20,000 facilities in the U.S. The research mission of the Department of Energy is carried out through grants to universities and other R&D providers but, most importantly, in the expansive federal laboratory facilities of the Department of Energy. The best known of these facilities are the "multi-program" program laboratories (often known as the "national labs"), all of which are owned by DOE but managed by contractors, either private firms or university consortia (or the two acting in concert). There are nine multi-program laboratories, the largest of which are Lawrence Livermore National Laboratory ($1,866,000,000 appropriation, 6,909 FTE), Sandia National Laboratories ($1,358,000,000, 7,500 FTE), and Los Alamos National Laboratory ($1,345,000,000, 6,942 FTE). But the DOE complex also includes "program-dedicated laboratories" (e.g. Ames Laboratory, Princeton Plasma Physics Laboratory) and "specific-mission laboratories" (e.g. Bettis Atomic Power Laboratory, Savannah River Technology Center).

Performance assessment activities of the Department center chiefly around annual budget submissions, Government Performance and Results Act requirements, and contracting and procurements standards. While the Department has made limited use of quantitative performance measures until recently, it has a long history of evaluation of its laboratory facilities, relying chiefly on peer review boards and site visit teams.

The organizational structure for the DOE is unusually complicated and its complexity affects performance assessment. The chief research providers of the DOE, the laboratory complex, have laboratory directors, but also report to DOE headquarters, regional offices (that vary in their degree of oversight), and managing contractors (such as Lockheed Martin and University of Chicago). A further complication is that most of the money for the laboratories comes from programs of the Department of Energy, especially the Office of Science (though DOE laboratories also compete for "work for others" funds, including funding from other government agencies, especially the Department of Defense). From the standpoint of research assessment, the key institutions are the DOE laboratories and facilities and the Office of Science, their chief source of funds.

Program structure. The formal mission of the Department of Energy is

> *"to foster a secure and reliable energy system that is environmentally and economically sustainable; to be a responsible steward of the nation's nuclear weapons; to clean up the Department's facilities; to lead in the physical sciences and advance the biological, environmental, and computational sciences; and to provide premier scientific instruments for the nation's research enterprise."*

Both the Department of Energy and its Office of Science have recently completed strategic plans. The Department identifies four "business lines," Energy Resources, National Nuclear Safety, Environmental Quality, and Science. The Plan includes goals, objectives and performance measures according to business lines. The general goal for Science is:

> *"to advance the basic research and instruments of science that are the foundations for DOE's applied missions, a base for U.S. technology innovation, and a source of remarkable insights into our physical and biological world and the nature of matter and energy."*

The Science goals includes four specific objectives (paraphrased here): (1) provide leadership in the physical sciences that will sustain the nation's energy needs; (2) develop scientific foundations to understand and protect the planet and support long-term environmental cleanup; (3) Explore matter and energy as elementary building blocks of life; (4) provide tools, scientific workforce, and research infrastructure to support the nation's scientific work.

Project assessment. Throughout the Department, the chief mechanism for assessing R&D projects is peer review. However, peer review varies considerably from one division of DOE to the next. In many respects, the Office of Science and, especially, the Basic Energy Science Division, have taken a lead in peer review of proposals for projects and, to a lesser extent, completed projects. The Office of Science peer review approaches have much in common with the procedures followed at the National Science Foundation. But other divisions of DOE have quite different approaches and procedures for peer review and have sometimes undergone criticism from the GAO and from expert panels.

The Office of Science provides traditional peer review (e.g. similar to processes for the National Science Foundation) for extramural research funds provided to universities and nonprofit organizations. Peer review generally is somewhat different when the proposal is from the federal laboratory system. In those cases the focus typically is on the program rather than individual projects. However, the Office of Science is committed to using peer review, even in those cases where proposals are from the federal laboratory system.

In the federal laboratory system there is considerable variation in approaches to project selection. In most laboratories division directories work with a group of senior scientists to allocate funds from programs to particular projects. What each of the multi-program laboratories has in common is the use of advisory

groups and site visit teams. Assessment of projects tends to be a managerial function carried out by division directors.

Assessment in the research process. During the research process, assessment mechanisms are generally informal and practices vary considerably among the federal laboratories. The chief assessment approach is informal assessment by management, especially the research and division directors.

The laboratories have two general approaches to assessing individual scientists. In some laboratories, scientists' evaluation is not greatly different from evaluation of other employees. An individualized performance plan is developed with specific objectives, targets and measures, often followed by a self-assessment. But at some of the laboratories (Brookhaven National Laboratory and Argonne National Laboratory) assessment procedures resemble closely those found in universities, even to the extent of having meaningful tenure reviews. In the laboratories employing a more academic style of personnel evaluation (i.e., the largest laboratories, *excepting* the weapons laboratories), the focus is on publications and conference presentations. These laboratories, as others throughout the system, have been slow to adopt broader criteria (e.g. working with industry, technology transfer).

Results and impacts. At the individual level, results and impacts are measured in concrete terms, often publications and papers produced. The focus is generally on output rather than impact. At broader levels, impacts are measured as part of strategic planning activities. In the Department's Annual Performance Plan, the measurement of results is, in recent years, based on concrete outcome indicators. Thus, for example, the Energy Resources 2001 Annual Performance Plan includes such specific and measurable targets as,

- Develop a 14% efficient stable prototype thin file photovoltaic module.
- Demonstrate carbon dioxide free production of hydrogen using a plasmatron at 30 kW scale.
- Complete testing and evaluation of a 5 MW Kalina Cycle demonstration geothermal power plant.

Such concrete measures were also presented in the 1999 performance plan and, in 2000, general assessment were provided (e.g., "exceeded goal," "nearly met goal").

Communication. The Department has taken strides to communicate extensively with the general public and also with specialized stakeholders. Much of the communication with the general public takes place via a quite comprehensive web site. Indeed, the web site includes so much material that it is not easy to filter information (especially with limitations of DOE search engines). The Department also communicates with the general public in town meetings and workshops, many pertaining to hazardous waste remediation at DOE sites. The meetings have often proved quite acrimonious but the Department is concerned about openness and transparency, in part because of a well-publicized history of inappropriate secrecy pertaining to its production of hazardous waste.

Pending changes. Currently, the Office of Science is in the process of radically restructuring their approaches to performance evaluation. The first part of this change is a three year study beginning with a comprehensive literature review (performed by Pacific Northwest National Laboratory) focusing on research studies pertaining to creativity, leadership, performance measurement and project management. This will be the centerpiece of a January 2001 workshop that will help the Office determine its performance evaluation needs. Other studies now under way examine the use (as performance measurement methods) of case studies, qualitative methods, data mining and benchmarking. As a result of these studies, the Office will refine performance measurement techniques and integrate them into the strategic planning and budgeting processes.

D. Environmental Protection Agency

The laboratories of the Environmental Protection Agency (EPA) are managed from the Office of Research and Development (ORD), which has an overall budget of about $550 million. Of this amount, about $150 million goes to a program of extramural grants, and about $400 million is spent in the laboratories. About 80 percent of the laboratory facilities are located in North Carolina, in the Research Triangle Park. Others are scattered around the country. The activities of this dispersed system are organized for planning purposes into three "national laboratories" and two "national centers."

EPA has thoroughly revamped its research activities over the last five years, in response to several hard external critiques. Scientific credibility had reached a low point, and a new system stressing peer review at many stages of the research process has been put in place. All extramural grants and contracts are now awarded through a competitive process, with peer review modeled on the National Science Foundation's processes. Agency-wide peer reviews for risk assessments are a central feature of the change. EPA has also taken the Government Performance and Results Act (GPRA) very seriously, making changes in its system of planning and priorities.

Program planning process. As called for in GPRA, EPA's program planning takes place in the context of the budget process. Each of 24 major areas in the research program has a long-range research strategy document, which has been subject to internal and external peer review through the agency's Science Advisory Board. Each area also has a National Program Director, charged with integrating the efforts within it. This person does not control resources, but plays a key role in communication among the units contributing to the area and between the area and others in ORD.

ORD is currently distilling each of these plans into a five-year research plan, which sets annual milestones in specific areas that indicate how the strategy will be implemented. These lay out the specific items each laboratory and center will produce in a given year. These implementation plans are in preparation, and will go through peer review and become public documents. They are designed to make EPA's plans and processes transparent to a variety of audiences. The new element in these documents is the five-year time horizon, which provides a more strategic context for the year-to-year decisions that are made in the annual budget process.

Priorities in the annual budget process. The annual budget process, however, allocates the resources to carry out these plans. To set the context for the annual process, the Assistant Administrator for Research consults with other Assistant Administrators for the agency on their priorities. Using this input, plus their knowledge of what is happening in the laboratories and extramural research, the senior ORD staff develop guidance for the next stage in the process.

Research Coordinating Teams (RCTs), established by theme, carry out this next phase. These teams include a representative from each laboratory and center, program officers from the regulatory offices, and ORD staff. ORD staffers bring knowledge of research directions. The regions bring information on issues that they are facing in various parts of the country. Discussion within the group results in a list of priorities, with the basic unit being larger than a project, but still fairly specific. For example, air pollution might have four or five sub-areas on the priority list. These sub-areas are ranked, with associated budget amounts, to total about 80 percent of the expected budget request. These are the items that will be funded, regardless of the results of the rest of the budget process. The RCTs also identify another 40 percent of the target budget, consisting of items that are desirable but at a lower level of priority. Some of these will end up being funded, depending on both agency and Congressional decisions.

The criteria for ranking the sub-areas are explicit and are spelled out in the EPA strategic plan. They focus sharply on impact. First, the RCTs ask, how much effect will this activity have on risk assessments? EPA is looking for a research base that is likely to result in large changes in risk assessments, rather than incremental tweaks. Second, how much will this knowledge affect risk management? EPA is looking for order of magnitude improvements in risk management strategies.

After the RCT priority-setting process, the senior staff of ORD looks across the categories in the resulting budget request to create a single ORD set of priorities. They report to an executive council composed of the Assistant Administrator for Research and the Center and Laboratory Directors. The request then goes forward through the normal budget process (agency; White House Office of Management and Budget; President's budget request to Congress; action by authorizing and appropriating committees).

Personnel and laboratory management. Early in the redesign process for EPA R&D, some wholesale changes were made at the laboratory level. But in general, the size and locations of laboratories is stable, in part due to local political pressures. As priorities shift through the program planning process, laboratory directors shift personnel from lower priority to higher priority projects. Personnel evaluation is left to division directors within the laboratories. The strong processes that have been established for peer review of the products of EPA research, however, are clearly an important influence on these evaluations.

Quality assurance for agency reports. Very well-defined procedures have been established for reports produced internally at EPA. Every laboratory has a Quality Assurance Manager, who examines data quality. Every project, internal or external, has a Quality Assurance plan. EPA results must be "golder than gold," according to the Assistant Administrator in charge of the system.

In addition, after the data has been incorporated into a report, it undergoes careful peer review. A Peer Review Advisory Group of senior administrators has been established. It has developed a peer review handbook aimed at managers, plus others who are using results. EPA's Science Advisory Board reviewed the document, as did the agency's Inspector General. A second edition is about to appear. Every major scientific or technical work produced must undergo peer review. Every Directorate nominates what will need review in the coming year, and needs to justify the decision not to review. A Quality Staff in the Office of Environmental Information maintains a database of the works to be reviewed, and audits what has been done. A key point for EPA is that the system for checking data and publication quality is transparent.

E. National Institute of Standards and Technology

The National Institute of Standards and Technology is an operating agency of the U.S. Department of Commerce and houses the central measurement and standards laboratories for the nation, an extramural R&D grant program, a manufacturing extension program, and the Baldrige National Quality Program. In FY 2000, the in-house laboratories at NIST received about $277 million in appropriations. Some laboratory operations were also supported through a Working Capital Fund of about $139 million; of that amount, $73 million was associated with research, development, and supporting services performed for other Federal agencies. In FY 2000, NIST employed 2740 people. The main facilities are located in Gaithersburg, Maryland, outside Washington, D.C., and there is also a facility in Boulder, Colorado.

Program assessment. Technical programs are evaluated both at the laboratory level and on a NIST-wide basis. Individual laboratories have their own evaluation processes, as each has a distinctive research portfolio and customer base. For the laboratories as a whole, program review is conducted centrally. The most extensive formal review is conducted annually by the National Research Council Board on Assessment of NIST Programs, which has been performing this peer review function since the 1950s. The NRC Board on Assessment organizes the review and is responsible for the final assessment; the peer review process itself centers on each laboratory and uses individual panels of industrial, academic, and other external experts that are selected by the NRC to match the competencies and focus of specific laboratories. The panels evaluate each laboratory independently, focusing on the technical quality of the work, program effectiveness, fit to stated needs, and whether the facilities are adequate. This process is the primary external check on quality and merit for the laboratory.

For example, in Fiscal Year 2000, the Board on Assessment of NIST Programs at the NRC consisted of six members, two from universities, two from industry, and two from independent research institutions. The overview section of their report noted that the overall technical merit of NIST programs remained high, their impact appeared strong, and mission relevance was good. The Board was worried about resource challenges, however. All seven NIST laboratories were reviewed, each by a separate panel chosen by the Board on Assessment. The Building and Fire Research Laboratory (BFRL), for example, was reviewed by a panel of nineteen members, of whom only seven were academics. The panel included representatives from the automotive, chemical, and

cement industries, along with private consultants, a lawyer, and an architect. For the laboratory overall and for each division, the reviewers commented on mission, technical merit, impact, and resources. In the Structures Division of BFRL, they judged that the current array of projects was aligned well with the mission. Comments on the technical merit of the research groups varied from "excellent progress" through "very good progress" to "satisfactory." The impact section of the review outlines dissemination audiences and mechanisms. The panel was quite worried about resources: "Laboratory equipment is in dire need of maintenance and upgrading."[9] Before the Board's report was issued, the NRC asked twenty-three outside reviewers to evaluate it, of whom six were from industry and two from other federal agencies.

NIST's Director and Deputy Director are also active in following the activities of the laboratories. They visit at least one laboratory a week, looking at fit to mission, work progress and outputs, external recognition, and relevance to customer needs. They communicate the results of their visits directly to laboratory directors, and point out any changes that are expected. They sometimes raise questions for the laboratory heads to address, for example, with regard to responding to new customer needs. These visits are used in part as a prodding tool.

The customer base for each laboratory is an important element of NIST's profile as a national service facility. The types of customers vary widely, however, from academic researchers to industrial customers in a wide variety of sectors. Given its diverse customer base, assessing customer needs is a challenge for the organization. Wherever possible, the laboratories participate in industrial technology road-mapping exercises or use other venues, such as trade and industrial associations, to bring diverse customer groups into the planning process and articulate metrology and standards needs.

Strengths and weaknesses. The weaknesses of the external review process at NIST are those that are commonly acknowledged among research evaluators. The peer review process is very expensive and time consuming. Staff time is used up at a high rate, as well the time of the reviewers. Furthermore, it is hard to communicate the results of such a process to external audiences, and it is very difficult to track trends over time or construct key performance indicators from the qualitative results of the process. It is hard to use this process to characterize the accumulation of accomplishments over time. The benefit that peer review delivers is the depth of information and technical detail it provides, which is particularly useful to front line managers.

The peer review approach also may not translate to all research organizations because of size issues. NIST can be comprehensive in its review process because it is relatively small. This strength of the process, comprehensiveness, is also related to its weakness, the difficulty in distilling results. It is a process that, if done well, should produce a sense of comfort and trust in oversight bodies, however.

Results and impacts. The examination of results is built into the review process just described. Expert review panels, however, can only guess at downstream impacts.

---

[9] Board on Assessment of NIST Programs, Commission on Physical Sciences, Mathematics, and Applications, National Research Council, An Assessment of the National Institute of Standards and Technology Measurement and Standards Laboratory, Fiscal Year 2000 (Washington, DC: National Academy Press, 2001).

Sometimes they try to judge potential impacts by comparing a program with similar types of efforts that already have been funded, using previous examples to ask whether they can connect current activities better to the intended impacts. Apart from external peer review, NIST uses two other evaluation mechanisms to assess results: 1) analysis of various output measures, such as publications and production figures for calibration services, reference materials, and the like, which need to be and are weighed differently in the evaluation of different laboratories; and 2) analysis of findings from microeconomic impact studies, which provide quantitative estimates of the downstream economic impacts from completed projects. While informative about ultimate impacts, the results from these economic studies cannot be translated into a measure of net value for the laboratories as a whole or any individual laboratory, especially not in the one fiscal year time frame that current accountability legislation calls for.

Communication of results. This is seen as a key challenge. Internally, the agency has developed a diversified measurement system: in addition to external peer review, it conducts economic studies of the impact of its programs and assesses various output measures ("signs of life") over time. Each evaluation method has its strengths and limitations, and it is difficult to convey concisely the net effect of the entire evaluation system. The economic impact studies have value in communicating, internally and externally, about how NIST programs generate results. For management, they increase understanding of the process of research and how it serves customers. These studies also provide management with detailed data on the timing and sequencing of impacts across different levels of industrial supply chains.

Externally, the signals are mixed about what evaluation method best communicates impact or "results". GPRA appears to prefer quantitative measurement, particularly outcome measurement. It is therefore useful for NIST to conduct and have evidence from the economic impact studies, even though they are not comprehensive but rather illustrative studies at the project level. NIST's measurement of social returns and cost-benefit studies are praised for being rigorous, but some external decision makers, including the Office of Management and Budget and the General Accounting Office, still come back for more information. They would prefer a net impact statement to a quantified case study. The most powerful tool for communicating the value of the programs to Congress, however, may be a narrative of program accomplishments in a given fiscal year in combination with top-level evidence from NIST's formal performance evaluation system. The information set needed for management is broader and more detailed than what is used for external reporting.

Benchmarking. Recent NIST Directors have expressed interest in comparing NIST's measurement capabilities at a technical level with those of other national measurement institutes. The exact differences are often quite difficult to confirm, and producing such data can be very labor intensive. They therefore do this selectively. Because the measure is one of capabilities, it does not provide much help with the problem of demonstrating outcomes of NIST's programs.

Changes in the last five years. Virtually everything mentioned here, with the exception of the international benchmarking studies, was part of NIST's repertoire before 1995. But more resources have been put into impact studies recently, in part because of GPRA. What GPRA has changed the most at NIST is the channels for communication. At first

NIST reported too many measures; they now select fewer to put forward in the GPRA performance plan and report. GPRA has also been useful in stimulating thinking about outputs and impacts.

F. National Institutes of Health

About ten percent of the almost $18 billion budget of the National Institutes of Health (NIH) is spent in its "intramural" program, located mostly at the NIH "main campus" in Bethesda, Maryland. The rest of the budget is distributed around the country through a highly competitive system of extramural grants to medical schools and universities. There are 25 Institutes and Centers (ICs) within NIH. Each basic science Laboratory or clinical Branch is integrated into the overall program thrusts of its IC. NIH is a major operating division of the Department of Health and Human Services (HHS), comprising over a third of its discretionary budget.

Program assessment. Research support is allocated to intramural scientists by their Scientific Directors and Laboratory/Branch Chiefs based largely on their demonstrated scientific accomplishments. Programmatic decisions within the intramural research programs rely heavily on the assessments of external Boards of Scientific Counselors (BSCs).

The Boards of Scientific Counselors provide evaluation and advice on the quality of science being done in the laboratory, including tenure-track candidates undergoing midterm or final review, and tenured Senior Investigators. The BSCs also address issues of resource allocation, specific projects including new areas of development, and other administrative matters. Their reviews are to evaluate the research program as a whole, for its overall goals, quality of research, and long-term objectives.

The BSCs have been in operation since 1956. In 1994, based on recommendations of an external advisory committee, NIH revised its policies and procedures for outside review and evaluation of intramural research by BSCs. The BSCs must be composed of individuals who have outstanding scientific credentials. The reports of BSCs are distributed to the NIH Director and Deputy Director for Intramural Research as well as the Scientific Director and Director of the Institute under review. The BSC also reports annually to the National Advisory Council or Board of the IC.

For each laboratory/branch being reviewed, the BSC receives the following information:

- A description of the overall past accomplishments of the Laboratory/Branch since the last review.
- A summary of the organizational structure of the laboratory reviewed.
- A listing of all personnel, including their position, type of appointment, and grade.
- Space usage.
- Operating budget; budget allocation procedures vary considerably among the Institutes and Centers.
- Outside contracts, if any.
- Cooperative Research and Development Agreements (CRADAs), if any.

Personnel assessment. Program review and personnel review are very closely tied at NIH. The number of intramural researchers at NIH has been stable for some time; the Bethesda campus is constrained by space. The NIH trains a large number of Postdoctoral Fellows (almost 3000) who may remain up to 5 years. New independent researchers are brought in through a system of tenure-track positions, which are advertised nationally and internationally. In addition to annual and ongoing internal performance reviews, each of these researchers receives a mid-term and final BSC review. At six (or eight if clinical or epidemiology) years, they have either received tenure, or left the Institute. After tenure, each researcher receives a careful review at least every four years by the Board of Scientific Counselors.

For each scientist being reviewed, the BSC receives the following information:

- A current CV and bibliography.
- Copies of up to three important recent manuscripts or publications.
- Details of ongoing research, including general aims of the research projects, overall past accomplishments since the last review, and a discussion of future research plans.
- A summary of the amount of support staff and space that the scientist uses, in addition to information about budget, contracts, and CRADAs.
- A listing of former fellows and their current positions.
- A copy of the most recent prior Board of Scientific Counselors report of the Laboratory/Branch under review is made available at the review.

Criteria for assessment. All reviews use the following criteria:

Significance. Have the investigator's studies addressed important problems? Are the aims of the project(s) being achieved? Is scientific knowledge being advanced, and are the projects affecting the concepts or methods that drive this field?

Approach: In general are the approaches well conceived? When problem areas arose, were reasonable alternative tactics used?

Innovation: Do the projects use novel concepts, approaches, or methods? Are the aims original and innovative? Do the projects challenge existing paradigms or develop new methodologies or technologies?

Environment: Is the investigator taking advantage of the special features of the NIH intramural scientific environment or employing useful collaborative arrangements?

Support: Is the support the investigator received appropriate?

Investigator Training: Is the investigator appropriately trained and well suited to carry out the projects being pursued? Is the work proposed appropriate to the experience level of the principal investigator and other researchers (if any)?

Productivity: Considering the investigator's other responsibilities (e.g., service or administrative), how would you rate his/her overall research productivity?

Mentoring: Is the investigator providing appropriate training and mentoring for more junior investigators?

Reporting of Results of Review. At the completion of the review, an oral summary of the review is given to the Scientific Director, Institute or Center Director, and Deputy

Director for Intramural Research (or their designees). In addition, the Board is encouraged to meet with the Laboratory/Branch Chief before adjournment. A written report is prepared following the format preferred by the Scientific Director after consultation with the Chair of the Board of Scientific Counselors. It consists of a narrative critique of the individual investigators and the research program of the Laboratory/Branch. The report is submitted to the Scientific Director. In Institutes and Centers that use site visit teams, the report is distributed to all members of the Board of Scientific Counselors. Recommendations of the site visit panel are considered by the entire Board, which then advises the Scientific Director.

Evaluations of individual investigators must address the quality of the research projects, the validity of the approaches used to address the scientific questions, and the level of resources (space, budget, and personnel) supplied to the investigator. These evaluations are written by members of the Board and reflect the majority view; minority views are also included. Each investigator receives his/her evaluation and has the opportunity to provide written comments to the Scientific Director.

A written report is sent within two months of the review to all Board members and ad hoc reviewers, the Scientific Director, and the Institute or Center Director.

Follow Up. At the next meeting of the Board, the Scientific Director responds to the report, indicating areas of agreement and disagreement and any possible actions. Within six months, the Scientific Director provides the Board with a written response. Copies of both the report and the response are sent to the Institute or Center Director, the Deputy Director for Intramural Research, and the Director, NIH. The Board of Scientific Counselors reports annually to the Institute or Center National Advisory Council, either by endorsing a written report of the Scientific Director, by providing the Board of Scientific Counselors report and Scientific Director's response, or by providing an independent report to be presented to the Council.

Results and impacts. The attention to results is a major difference between quality control in the intramural and extramural programs. Selection of extramural projects is based on prospective plans, but the bulk of review of intramural programs and people is based on track record. There is, however an increased emphasis toward taking future research plans into account in intramural review as well.

G. Naval Research Laboratory

The Naval Research Laboratory (NRL) is the main research branch of the U.S. Navy. It is funded on a scheme that is close to unique among U.S. government laboratories: a working capital fund. This means that NRL does not receive its own appropriation through the regular government budget process, but rather "sells" its services to a variety of government customers. Thus, a satisfied customer is the key metric of success for the organization.

The total cost of keeping the Laboratory running is about $800 million. This budget pays NRL employees, the overhead, and a set of contract partners who also participate in the research. (About half the projects of the labs involve external partners.) The Laboratory

employs about 3000 people, including 1500 scientists and engineers, and has another 1500 or so de facto employees at the partner organizations.

The funding and management system is strongly project-based. Allocation of resources across the 18 NRL divisions results from the kind of work they are able to sell to their customers and projects that are chosen as part of what is called the Base Program for the Laboratory. The work stretches from basic research through testing and development, across the Department of Defense RDTE categories (Research, Development, Test, and Evaluation).

Base. Program About 20 percent of the Laboratory's funding comes from the Office of Naval Research, the overall coordinating office for all of the Navy's research activities, through what is called the Base Program. The Laboratory's Research Advisory Committee (RAC) plays a key role in developing this proposal and in allocating funds from the Base Program. The RAC is chaired by the NRL Director of Research, and includes the Associate Directors, who head NRL's three major research areas: ocean and atmospheric science and technology, materials science and component technology, and systems (including radar and information technology, for example). Seven "focus area coordinators" (FACs), who keep track of technical areas across the Laboratory, also participate.

What they propose to ONR draws on their deep and regular contact with what the operating divisions are doing for customers and what they need to maintain their capabilities. The RAC uses the Base Program to make investments in new areas or to deepen core competencies. The projects funded from this Program tend be more toward the basic research end of the spectrum. Projects are chosen through a competitive process. The RAC provides the divisions with a vision of where they think things are going in NRL's area, but the process allows bottom-up input. A compelling new development could emerge and be funded through this route.

The competitive process starts within divisions. Each division receives a "turnover" target from the RAC -- the percent of its funds that should be devoted to new activities that year. It is expected to send proposals totaling twice that amount to the next stage of the selection process. Proposers may also send proposals directly to the RAC, if they think they have a good idea that is being squeezed out by division priorities. The RAC members rate each proposal using a ten-point scale on several criteria. These include science and technology merit, the credentials of the performers, facilities available, Navy relevance, and transition potential (the likelihood that the work will be used somewhere, whether that is in the Navy, some other part of the military, or the private sector). The raters also give an overall score.

As with many peer review processes, this scoring sorts the 90-100 proposals received each year into those that are clearly at the top and should be funded, those that are clearly at the bottom, and a group in between. The RAC tries to meet the required turnover with high scorers. Division Directors have the opportunity to fund projects that do not receive RAC approval by shutting off something they already have going in their laboratories, under previously approved Base Program funding.

The projects supported through the Base Program go through formal retrospective program review, with a third of the projects up for review each year. These reviews are

by external teams, who rate the projects on criteria that are similar to those used in project selection. Copies of the reports of these groups go to the Director of ONR.

Customer supported programs. The bulk of the Laboratory's work, about 80 percent, is evaluated directly by the specific customers who sponsor it, in the process of deciding whether to continue, expand, or shrink support. These customers include the Navy systems commands; the Army, Air Force, and Defense Advanced Research Projects Agency (DARPA); and other government agencies, including the National Aeronautics and Space Administration (NASA) and the Department of Energy (DOE). Sometimes units of NRL are actually co-located with the customer organization, symbolizing the integrated working relationship. In this kind of relationship, track record plays a critical role, so results are constantly being taken into account in the continued success of the division or laboratory.

Personnel assessment. The Laboratory follows Standard personnel assessment practices with its government employees. The Laboratory tracks publications, citations, patents, and CRADAs (cooperative research and development agreements), and uses these in personnel decisions. NRL's basic researchers are encouraged to stay active in their technical communities, and go through regular peer review.

A striking feature of the ONR project-based system is that it provides a clear criterion for identifying employees whose work is not finding sponsors and who should be let go. Perhaps as much as three percent of NRL's government employees leave the Laboratory through this route every year. Combined with the heavy use of contract partners, this system provides a high level of personnel, and therefore program, flexibility.

Strengths and weaknesses. One strength of the project-based system is its pluralism. If a laboratory depends entirely on a single appropriation, the spigot can be turned off. This would be quite unlikely in the NRL situation. A weakness is that it can result in "frenetic marketing" -- the loss of time and energy to finding project work. Occasionally, a laboratory's customer base will thin and it begins to "take in laundry" -- that is, do routine work that does not meet the agency's mission just to meet the payroll. At NRL, such a situation is apparent to senior managers immediately. The project system provides the flexibility to shut down a laboratory that has adopted "laundry" as a way of life. This has happened recently.

Recent changes. The working capital fund system has been in placed since the 1950s. The Base Program process has been developed since 1995.

Interviews

Dr. Philip Chen, Senior Advisor to the Director of the Division of Intramural Programs, National Institutes of Health

Dr. Timothy Coffey, Civilian Director of Research, Naval Research Laboratory

Dr. Paul Doremus, Office of Strategic Planning and Economic Analysis, National Institute of Standards and Technology

Dr. Norine Noonan, Assistant Administrator for the Office of Research and Development, Environmental Protection Agency

Dr. Peter Preuss, Director, National Center for Environmental Research, Environmental Protection Agency

Dr. David A. Rust, National Program Staff, Agricultural Research Service

Dr. William Valdez, Director, Office of Planning and Analysis, Office of Science, Department of Energy.

Bibliography

Agency sources:

ARL web site: http://www.arl.army.mil/.

ARS web site: http://www.nps.ars.usda.gov.

DOE Office of Science, Office of Planning and Analysis web site: http://www.sc.doe.gov/sc-5/..

EPA/ORD web site: http://www.epa.gov/ORD.

NIH Office of Intramural Research web site: http://www1.od.nih.gov/oir/sourcebook/oir/oir-staff.htm#Overview.

NIST web site: http://www.nist.gov/public_affairs/guide/index.htm

NRL web site: http://www.nrl.navy.mil/index.htm

Other reading:

Brown, Edward A.1996. "Conforming the Government R&D Function with the Requirements of the Government Performance and Results Act. Planning the unplannable? Measuring the unmeasurable?" Scientometrics 36:3.

Committee on Science, Technology, and Public Policy, NAS/NAE/IOM. 1999. Evaluating Federal Research Programs: Research and the Government Performance and Results Act Washington, DC: National Academy Press.

Cozzens, Susan. 1999. "Are New Accountability Rules Bad for Science?" Issues in Science and Technology, .Summer, pp. 59-66.

Crow, Michael, and Barry Bozeman. 1998. Limited by Design: R&D Laboratories in the U.S. Innovation System. New York: Columbia University Press.

Kostoff, Ronald. "GPRA Science and Technology Peer Review." Available at http://www.sciquest.com/cgi-bin/ncommerce3/ExecMacro/sci_kostoff2.d2w/report

National Academy of Sciences, National Academy of Engineering, Institute of Medicine. 1995. On Being a Scientist: Responsible Conduct in Research, Second Edition. Washington , DC: National Academy Press